DRAFT 2016-07-29

Soc 650/Stat 503: Models for Categorical Outcomes

Spring 2016

Professor:	Scott Long
Lectures:	Tu & Th 1:00pm – 2:15pm Ballantine 305
Office Hours:	Tu & Th 11:00am – 2:00pm Ballantine 842b (across from elevators) If I am talking with someone, let me know that you are waiting.
Assistants	Bianca Manago (bmanago@umail.iu.edu); Roshan Pandian (rpandian@indiana.edu)
E-mail:	jslong@indiana.edu (please put "CDAiu: <topic>" in the subject line)

Overview

Categorical Data Analysis is a course in in applied statistics that deals with regression models in which the dependent variable is binary, nominal, ordinal, or count. The models studied in the class include probit and logit for binary outcomes, ordered logit and ordered probit for ordinal outcomes, multinomial logit for nominal outcomes, and Poisson, negative binomial, and zero inflated models for counts. Other less common models are considered briefly. Basic algebra, probability theory, and concepts from calculus are used to explain the structure and assumptions of each model. These ideas are used to develop sophisticated methods of interpretation that deal with complications introduced by the nonlinearity of the models. The course assumes that you have had a prior class on linear regression.

Books

1. Lecture notes are available as PDF files on Box.
2. Long, J. Scott and Freese, Jeremy. 2014. *Regression Models for Categorical Dependent Variables Using Stata, 3rd Edition*. Stata Press: College Stata, TX is required.
3. Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage is not required but is useful for those who are interested in mathematical details.
4. Long, J. Scott. 2008, *The Workflow of Data Analysis Using Stata*. Stata Press: College Station, TX. If you plan to do a lot of data analysis, I highly recommend this book.

Policies

1. Arrive on time and be ready for lecture at 1:00pm.
2. You will receive an invitation to the Box directories containing materials for the class and where you will submit your work. If you do not received an invitation, please send me an email from you IU e-mail account with subject line CDAiu: Box invite.
3. Lecture slides and other handouts are available on Box. You are welcome to print these.
4. Computing labs are extremely important. While most of lab is for you to work on your assignments, the TAs will sometimes give short presentations or discuss assignments at the start of lab. TAs are available for the first 90 minutes of lab, but might not be available in lab the last 30 minutes.
5. The class uses Stata 12 or later. Some assignments might require Stata 13. Computing clusters on campus and IUanywhere have Stata 14. You can purchase a student license; for details, google "Stata grad plan".

6. **Assignments are due at the start of lab on the day they are due!** One late assignment is accepted without penalty; it is due at the beginning of the next lab. Additional late assignments are penalized 25% if they are turned in no later than the start of the next lab. Exceptions are made for special circumstances after contacting the instructor after class (please not before class!) or during office hours. If we agree to an extension, you must send me an e-mail with our understanding and I will reply with a confirmation.
7. You are encouraged to share what you learn with others and help one another with problems. What happens in your model might be quite different from what happens in some other model. So, please show each other what you find. Of course, you must do your own work and you cannot use the same model specification used by other students.
8. To get to my office, enter the door 842 without knocking. My office is at the end of the hallway. If I am talking with someone, let me know you are waiting. Feel free to contact me by e-mail; during the week if you don't hear within 12 hours, try again.

Assignments and grading

There are 11 major assignments and there might be a few smaller assignments given in class:

1. Assignment 1: Math review. Everyone will repeat this assignment till they get a perfect score.
2. Assignment 2: Data cleaning.
3. Assignment 3: Picking your variables.
4. Assignment 4: Linear regression models.
5. Assignment 5: Binary regression models – part 1.
6. Assignment 6: Binary regression models – part 2.
7. Assignment 7: Testing and Fit.
8. Assignment 8: Nominal regression models – part 1.
9. Assignment 9: Nominal regression models – part 2.
10. Assignment 10: Ordinal regression models.
11. Assignment 11: Count models.

Due dates for the assignments are given in \Box\CDAstudents\+ Assignment submission\. Due dates might be changed if we get behind in lecture or students encounter problems. Handouts for the assignments are located in \Box\CDAresources\Assignment handouts\. You will edit these docx files to include your answers. Completed assignments are due at the start of your lab on the due date with associated files posted at \Box\CDAstudents\ + Assignment submission\.

Grades are based on assignments, your workflow, and attendance. Each assignment is given a number of points with a total of about 1000. Your grade is based on your percent of the total points using A=100-94%; A-=93-91%; B+=90-88%=B+; B=87-84%=B; B-=83-81%; 80-71%=C; 70-61=D. The highest grade possible if your script files do not reproduce the results from your assignments is a B (see Workflow below).

I apologize if a mistake is made in grading. Return the assignment to me at the end of lecture along with a cover page explaining the error. If I do not return the assignment documenting the change within two class periods, remind me by e-mail. Multiple people are grading and we try very hard to be consistent.

Workflow requirements for assignments

An essential part of data analysis is being able to reproduce your findings. Many journals are requiring that you submit your data and script files before a paper is accepted; if the paper's results cannot be reproduced, the paper is rejected. If the paper is accepted and others find errors in your analysis, a retraction is possible.

Since this is an applied class in data analysis, a portion of your grade is based on the workflow used to complete your assignments. A handout on workflow is presented and discussed in lab.

More general information on a reproducible workflow is in Long's *The Workflow of Data Analysis Using Stata*. For this class you are not required to implement the full workflow from the book, but you will be required to improve your workflow to allow reproducible results. I encourage you to attend my WIM talk Reproducible Results and the Workflow of Data Analysis on August 26 from 1-2:30 in Social Science Research Commons Grand Hall (Woodburn Hall 200).

Math and Stata

1. Stata is used for assignments along with the SPost13 package for Stata written by Long and Freese. If you haven't used Stata, I encourage you to look at the Stata channel on YouTube. On Box you will also find some materials on *Getting Started with Stata*.
2. The course assumes you are comfortable with basic algebra and elementary probability. If you are rusty, the first assignment should help. Additional math review is on Box in *cdaiu math review 2007.pdf*.

Getting help

1. If you need help debugging a program, the best thing is to place relevant files in the Box directory \-Help me\ within a subdirectory for your problem. See the example folder for what you should do. Include the do-file, the dataset, and log file in text format, not smcl.
2. The do-file must be self-contained. It must load the data, create needed variables (if any), generate the problem, and save a log file in text format. The do-file must have comments explaining what you are doing and what the problem is.
3. If a SPost command is causing a problem, include the command `which` *command-name* for the command causing the problem. This tells me which version of the command you are using.
4. Do **not** refer to specific directories (e.g., do not: use d:\mydata\science3.dta). Assume that your data is located in your Stata working directory.

Here is an example of what the do-file might look:

```
capture log close
log using jslong-a04-problem, text replace

// Scott Long – 2016-07-27
// Assignment 4: binary regression
// ERROR: see #3 below.

// #1: load data and check data
spex science2, clear
tab y
sum x1 x2

// #2: estimate logit
logit y xl x2, nolog

// #3: compute discrete change
// ERROR: variable xl not found
```

```
which mchange
mchange, at(x1=1 x2=3)

log close
exit
```

5.

cdaiu16 syllabus 2016-07-29.dox